# Data Exploration and Analysis: Three Algorithms

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2017
Columbia University

# Housekeeping

- Data challenge due today 6PM
- Today:
  - your graphs
  - continue with visualization and analysis
  - Team progress report
- next week: your second progress report...

# Your best graph ever!

# Data challenge 1: a recap

# Data challenge 1

## a quick review...

*86.1% of dead civilians who presumably participated in confrontations with federal armed forces were killed in events of "perfect lethality" where there were only dead and no wounded. [...] Mexico has the terrible situation of having lethality indices of 2.6. The lethality index of the Federal Police is 2.6 dead for every wounded, the Navy's reaches 17.3 dead for every wounded, and the Army's is 9.1 dead for every wounded.*

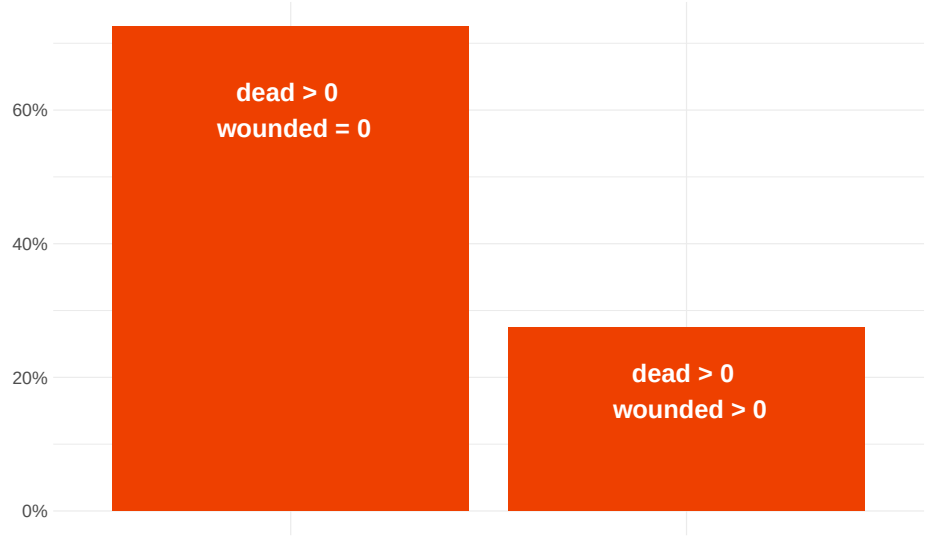
1. **Can you replicate the 86.1% number? the overall lethality ratio? the ratios for the Federal Police, Navy and Army?**

# Data challenge 1

## a quick review...

First, based on the database, we can compute that:

- **78% of all organized crime deaths happened in events of “perfect lethality”**




- ... we were aiming to find **86.1 %**

# Data challenge 1

## a quick review...

Second, we need to calculate a lethality indices:

- How do you compute a lethality index?
  - i) the ratio of **total** deaths over **total** wounded among organized crime

$$\frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

  - ii) the average of the **individual** ratios computed at the event level

$$\frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{w_i} \tag{2}$$

- what is the **substantive difference?**

# Data challenge 1

## a quick review...

- And there is also a **numerical difference!**

| | **overall** | **army** | **navy** | **federal police** |
|---|---|---|---|---|
| index (original) | 2.6 | 9.1 | 17.3 | 2.6 |
| index (total) | 3.0 | 5.4 | 4.6 | 3.0 |
| index (avg) | 0.74 | 0.63 | 0.68 | 0.45 |

# Data challenge 1

## a quick review...

- There is a difference because the numbers were quoted from a study unrelated to this data!

**Tabla 5.** Índice de letalidad de presuntos delincuentes fallecidos sobre presuntos delincuentes heridos

| | |
|---|---|
| Policía Federal | 2.6 |
| Ejército | 9.1 |
| Marina | 17.3 |
| Policía Federal y Ejército | 4.8 |
| Fuerzas de seguridad | 7.3 |

Fuente: Base de datos de enfrentamientos (prensa, enero 2008-mayo 2011).

Figure: Silva et al. (2012)

# Data challenge 1

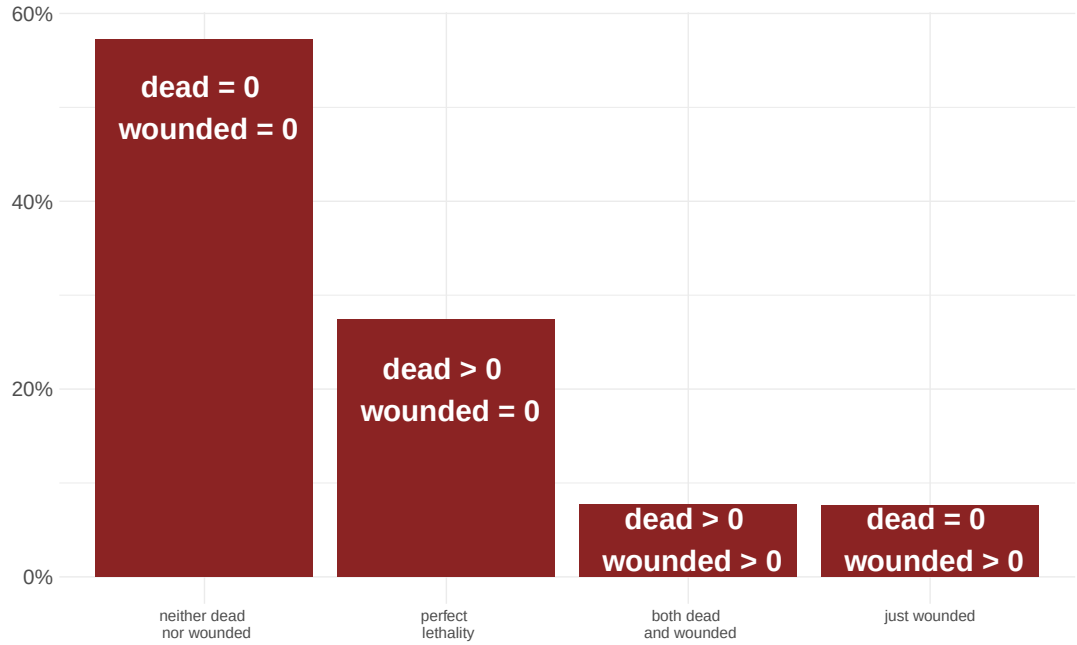
## a quick review...

2. The additional questions:
   - Is this the right metric to look at? Why or why not?
   - What is the "lethality index" showing explicitly? What is it not showing? What is the definition assuming?
   - With the same available data, can you think of an alternative way to capture the same construct? Is it "better"?
   - What additional information would you need to better understand the data?
   - What additional information could help you better capture the construct behind the "lethality index"

# Data challenge 1

## a few steps forward...

- Is there more in the data that we may be missing?

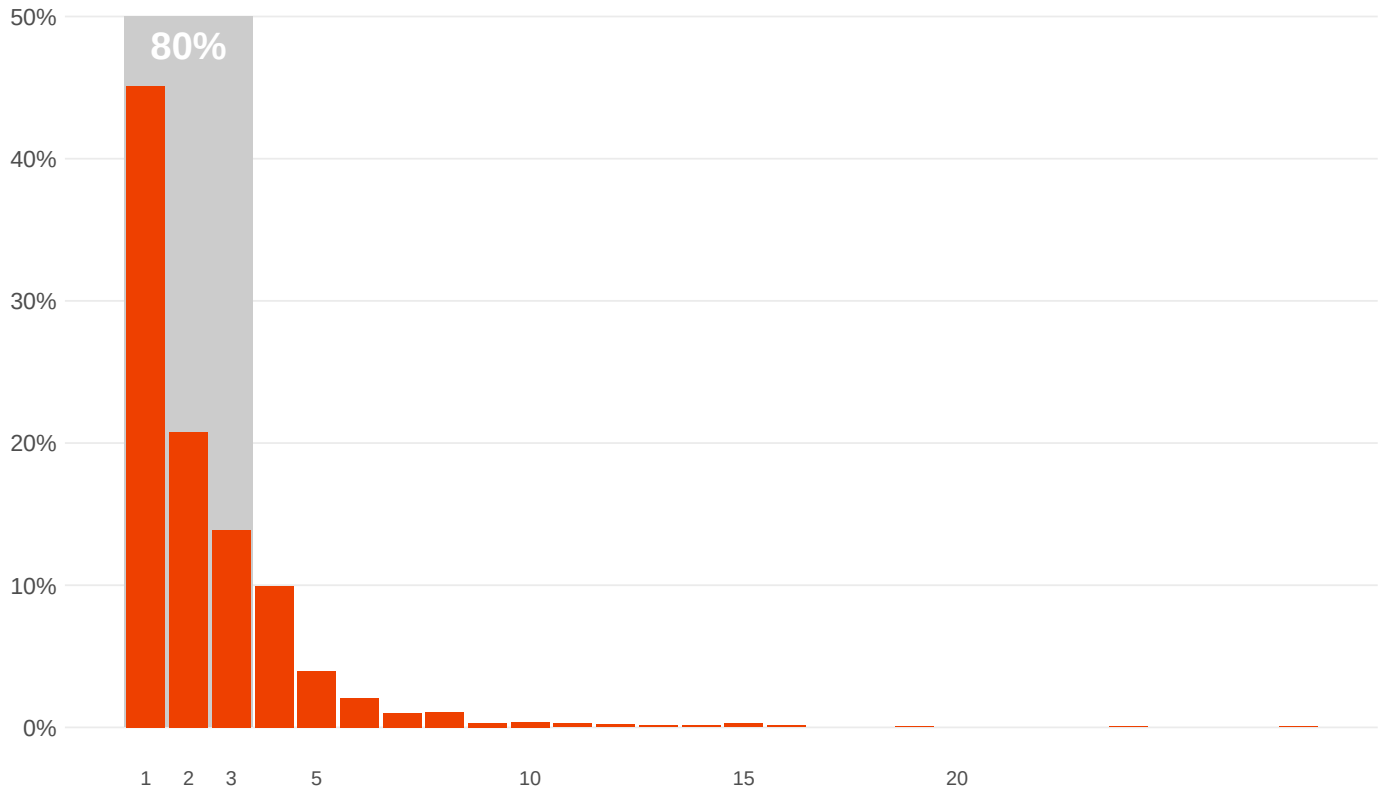



- the full database contains **5,396** events
  - in **57%** of events there were **neither deaths nor wounded**
  - in **27%** of events there is **perfect lethality**
  - in **8%** of events there were **both dead and wounded**
  - in **8%** of events there were **just wounded**

# Data challenge 1

## a few steps forward...

- 80% of events of “perfect lethality” had between 1 and 3 deaths



# Data challenge 1

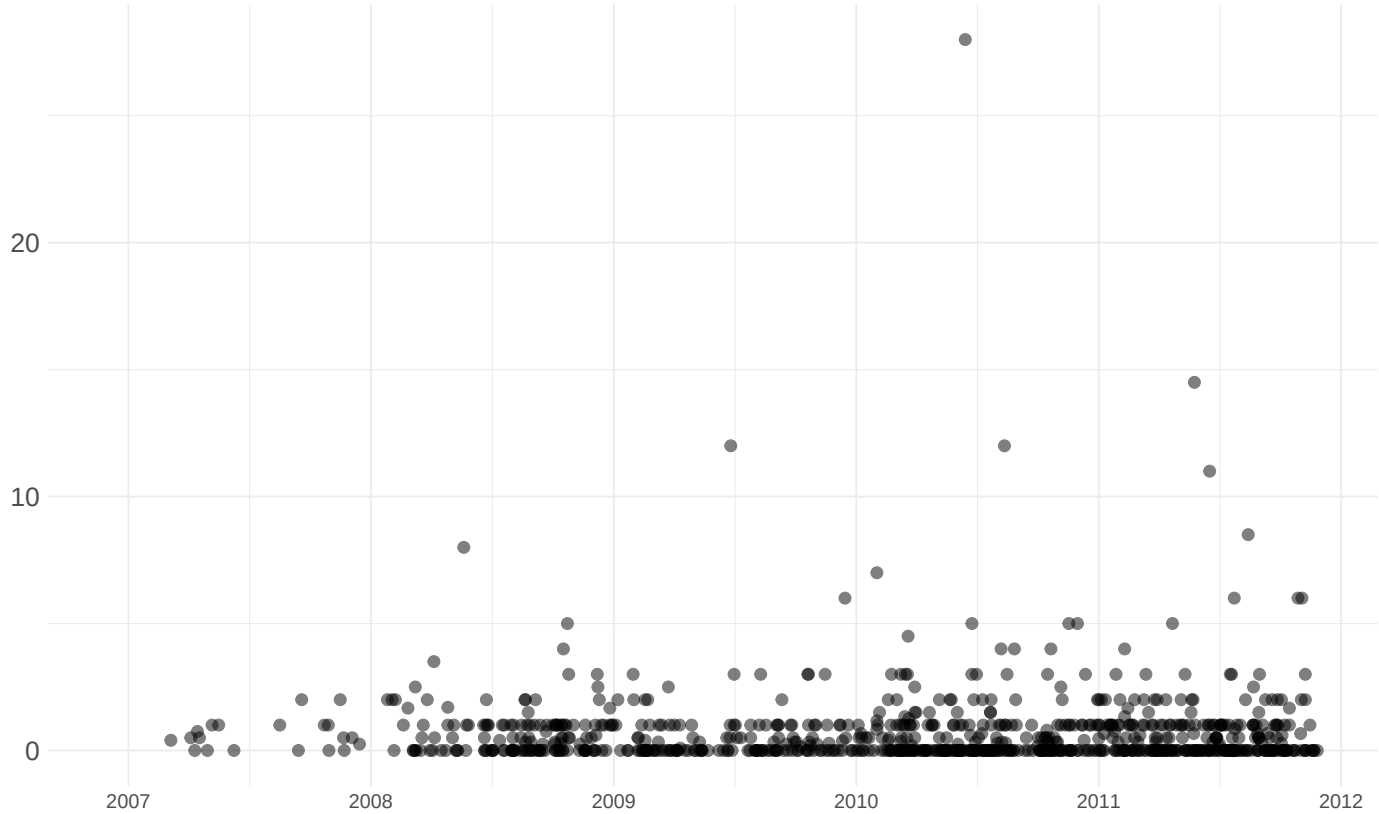
## a few steps forward...

- All summary statistics have pros and cons
- Lethality index shows the proportionality between dead and wounded
- some edge cases
  - **all dead, no wounded** is excluded since the ratio is undefined (e.g. $\frac{8}{0}$ is undefined)
  - **no dead, no wounded** is also excluded since the ratio is again undefined (e.g. $\frac{0}{0}$ is undefined)
  - **no dead, all wounded** is misleading as it gives the same value whether it was one or a thousand wounded

# Data challenge 1

## a few steps forward...

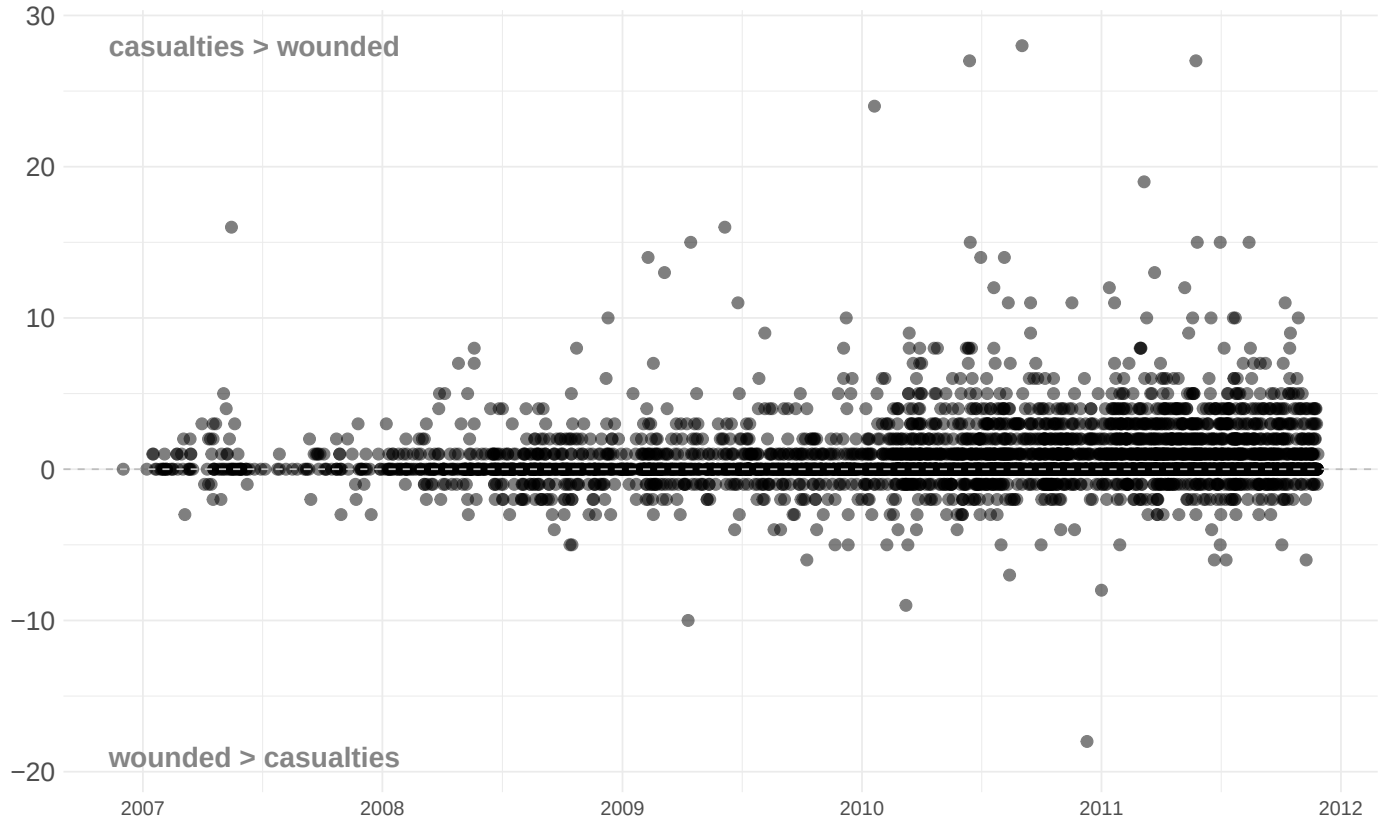
- the lethality index can only be computed for **825 cases (16% of events)**



# Data challenge 1

## a few steps forward...

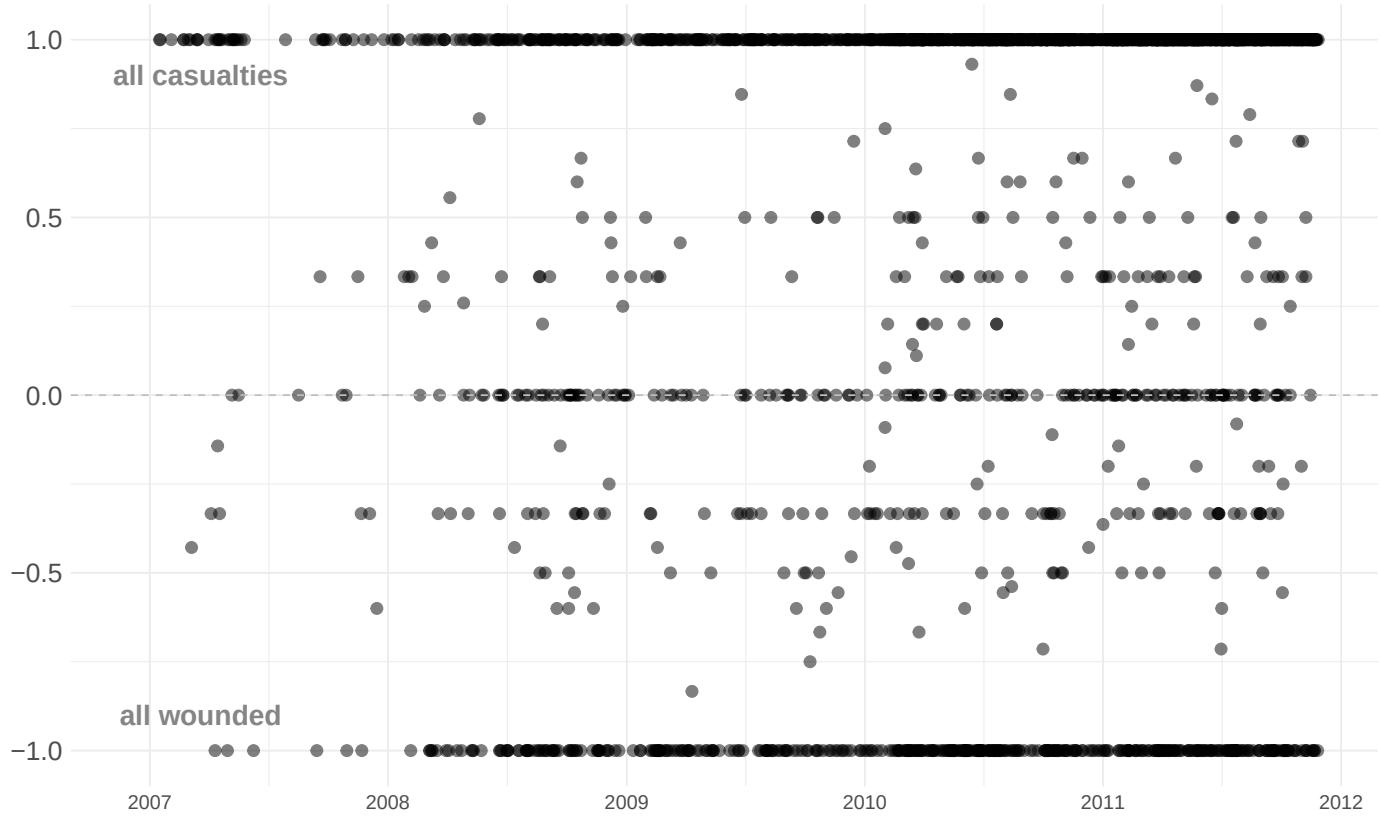
- if we're interested in the relation between dead and wounded, a simple difference may be illustrative: $d_i - w_i$



# Data challenge 1

## a few steps forward...

- with such variable range, perhaps we need to normalize:
  $\frac{d_i - w_i}{d_i + w_i}$




- what did we lose with the normalization? what did we gain?

# Three Algorithms

## Algorithm 1: OLS

- what is a regression?

$$E[y|x] = f(x)$$

- where $f(x)$ is a conditional mean function, such that

$$y = E[y|x] + \epsilon$$

- empirically: what do we get from a regression?

# Three Algorithms

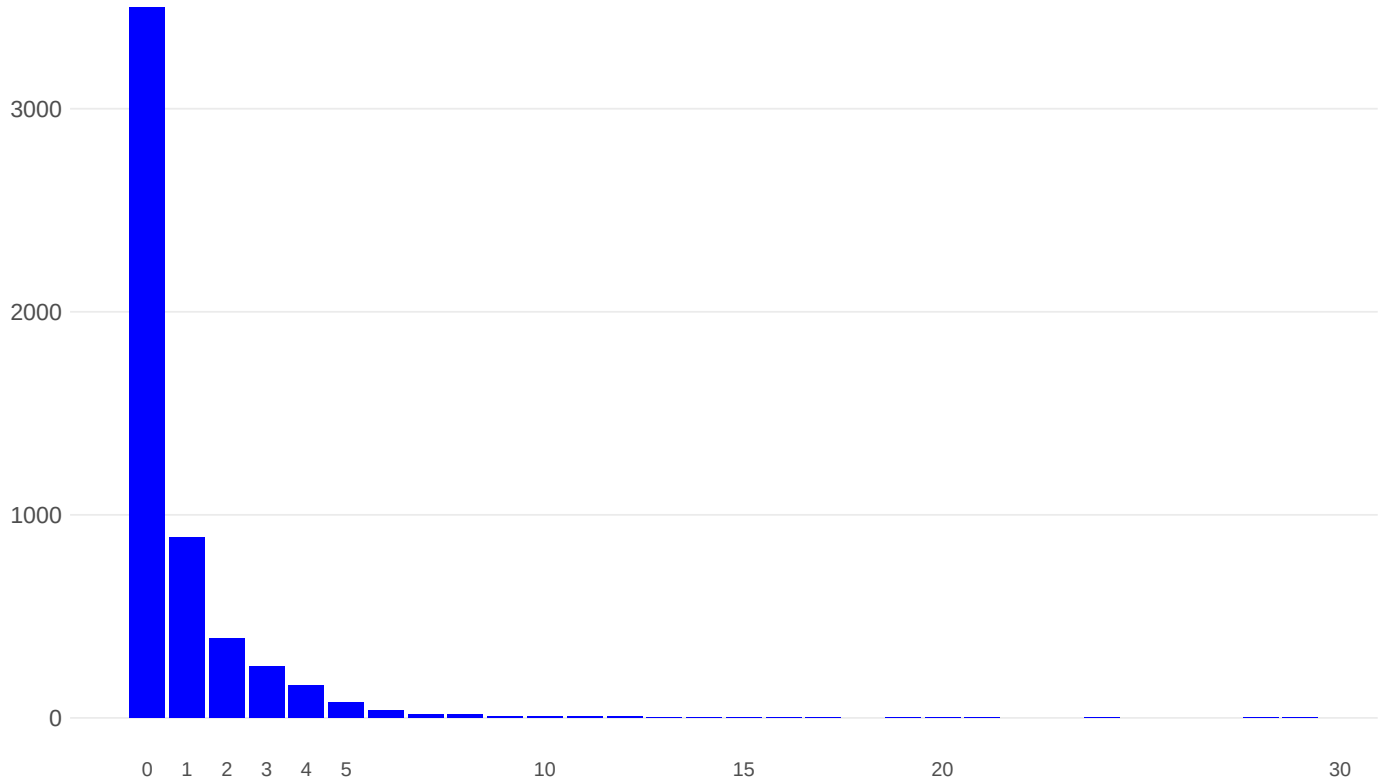
## Algorithm 1: OLS

```
Call:
lm(formula = organized.crime.dead ~ organized.crime.wounded +
    afi + army + navy + federal.police + long.guns.seized + small.arms.seized +
    clips.seized + cartridge.sezied, data = AllData)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6058  -0.7274  -0.4506   0.2192  27.3262

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.4505553  0.0332307  13.558  < 2e-16 ***
organized.crime.wounded  0.3736900  0.0239171  15.624  < 2e-16 ***
afi                     -0.2261752  0.4210396  -0.537   0.5912
army                     0.3066898  0.0532594   5.758 8.96e-09 ***
navy                     0.7150402  0.1389449   5.146 2.75e-07 ***
federal.police          -0.1271515  0.0773309  -1.644   0.1002
long.guns.seized         0.1478424  0.0085972  17.197  < 2e-16 ***
small.arms.seized       -0.0437447  0.0184592  -2.370   0.0178 *
clips.seized             0.0004374  0.0003152   1.388   0.1653
cartridge.sezied        -0.0001690  0.0000193  -8.760  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1  ‘ ’ 1

Residual standard error: 1.731 on 5386 degrees of freedom
Multiple R-squared:  0.1413,Adjusted R-squared:  0.1398
F-statistic: 98.44 on 9 and 5386 DF,  p-value: < 2.2e-16
```

# Three Algorithms

## Algorithm 1: OLS

- but wait... what does my DV look like?




- we can always log it, right? think again...

# Three Algorithms

## Algorithm 1: OLS

- assume for a moment that this is not problematic
  - (it is! but assume..)
- when analyzing people and behaviors, we're not only concerned about levels
  - we typically care about behaviors *conditional* on something else happening
  - note that this is different from "holding the rest constant"
  - these can be easily computed through **multiplicative interactions**

# Three Algorithms

## Algorithm 1: OLS

- from the model (with multiplicative interactions)

$$Y = \beta_0 + \beta_X \mathbf{X} + \beta_Z \mathbf{Z} + \beta_{XZ} \mathbf{XZ} + \epsilon$$

- we'd be interested in the marginal effect

$$\frac{\partial Y}{\partial \mathbf{X}} = \beta_X + \beta_{XZ} \mathbf{Z}$$

# Three Algorithms

## Algorithm 1: OLS

- Always, always, always remember:
  1. Use multiplicative interaction models **whenever one's hypothesis is conditional** in nature.
  2. Include **all constitutive terms** in the model specification.
  3. **Do not interpret the coefficients on constitutive terms as if they are unconditional marginal effects.**
  4. Do not forget to **calculate substantively meaningful marginal effects and standard errors**.
- ... or face the wrath of econometricians

# Data Exploration and Analysis: Three Algorithms

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2017
Columbia University